STA 35C: Statistical Data Science III

Lecture 17: Regularization Methods (cont'd) & Multiple Testing

Dogyoon Song

Spring 2025, UC Davis

Announcement

Midterm 2 on Fri, May 16 (12:10 pm–1:00 pm in class)

- **Arrive early**: The exam starts at 12:10 pm and ends at 1:00 pm sharp
- **One hand-written cheat sheet:** Letter-size (8.5"$\times$11"), double-sided, brief formulas/notes
- **Calculator**: A simple (non-graphing) scientific calculator is allowed
- **No other materials** beyond the single cheat sheet (no textbooks, etc.)
- **SDC accommodations:** Confirm scheduling with AES online ASAP

Preparation tips:

- Primary coverage: Lectures 12–19 (including next Wed)
- Key concepts from earlier topics may be assumed (cf. Midterm 1 Problems 2-4; HW 3 Problems 1-3)
- A practice midterm and brief solution key will be posted on course webpage
- Office hours next week:
    - Instructor: Wed, 4–6pm (extended); no OH on Thu
    - TA: Mon/Thu 1–2pm

Today's topics

- **Regularization**: More details
  - Recap: Ridge vs. lasso
  - Closer look into the shrinkage effects
  - Geometric intuition
  - Comparison of ridge vs. lasso
- **Multiple hypothesis testing**: Motivation
  - Why single-hypothesis testing may fail in large-scale settings
  - Type-I error inflation and how to control it

Recap: Why regularization?

Challenges: Least squares estimates...

- Can be unstable or undefined when $p \approx n$ or $p > n$, or if data are noisy
- May fail to capture a "sparse" underlying relationship

Regularization can stabilize estimation by adding a penalty term: with $\lambda \geq 0$,

$$\hat{\beta}_\lambda \in \arg \min_{(\beta_0, \beta_1, \ldots, \beta_p)} \Bigg\{ \underbrace{\sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2}_{\text{RSS}} + \lambda \underbrace{R(\beta_1, \ldots, \beta_p)}_{\text{penalty}} \Bigg\}$$

- The penalty shrinks coefficients to reduce variance at the cost of some bias

Two popular choices:

- **Ridge**: $R(\beta_1, \ldots, \beta_p) = \sum_{j=1}^{p} \beta_j^2$
- **Lasso**: $R(\beta_1, \ldots, \beta_p) = \sum_{j=1}^{p} |\beta_j|$

Ridge: Regularization reduces variance with shrinkage

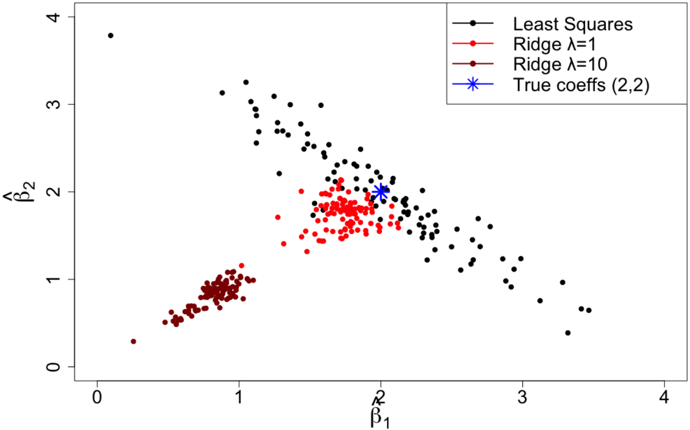

Figure: Scatter plots of 100 least squares estimates (**black**) vs. ridge estimates for $\lambda = 1$ in **red** and $\lambda = 10$ in **dark red**. As λ grows, the estimates cluster more tightly (lower variance) but shift away from the true value (blue star), indicating increased bias.

Ridge: Contours of training objective functions

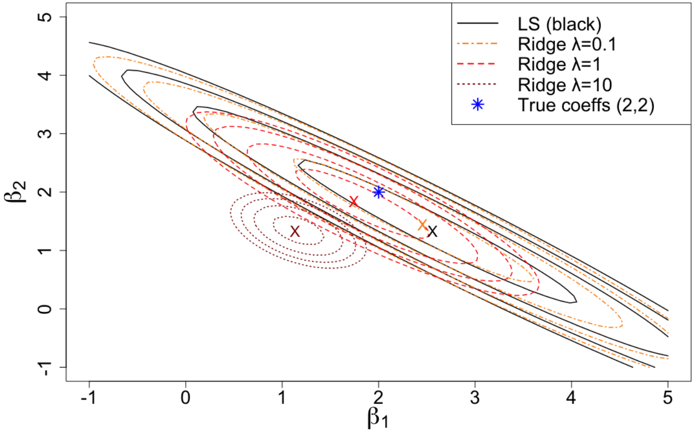

Figure: Contour plots of the least squares objective function (=RSS) in **black**, ridge regression objective for $\lambda = 0.1$ in **orange**, $\lambda = 1$ in **red**, $\lambda = 10$ in **dark red**. As λ increases, the ridge minimizer moves closer to $\beta = 0$. This depicts a single instance of data.

Ridge: Illustration with 1D example

In the simplified setting with $n = p = 1$ without intercept, ridge solves for $\lambda \geq 0$:

$$\hat{\beta}_\lambda^R \in \arg\min \left\{ (y - x\beta)^2 + \lambda\beta^2 \right\}$$

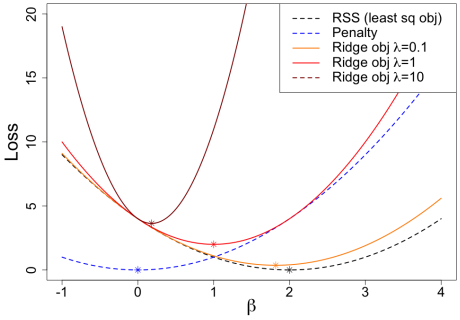

Figure: As λ grows, $\hat{\beta}_\lambda^R$ shrinks toward 0 for fixed (y, x) ($y = 2$, $x = 1$).

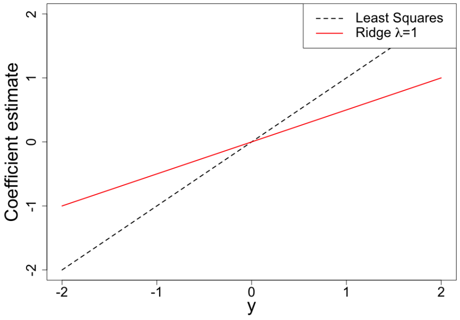

Figure: For each y, $\hat{\beta}_\lambda^R$ is smaller than the LS estimate y/x in magnitude, when $\lambda > 0$.

Lasso: Regularization reduces variance, but...

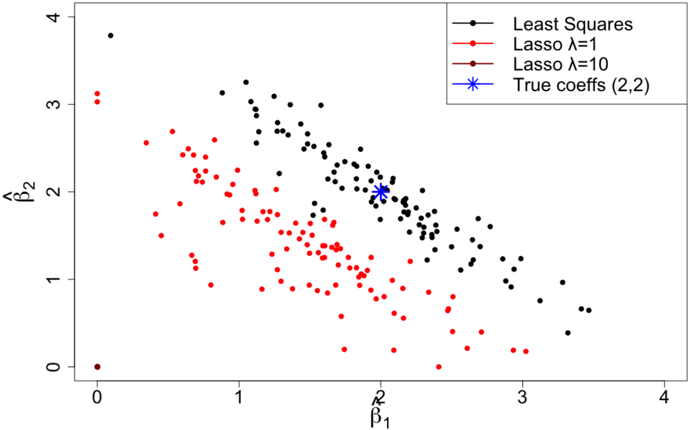

Figure: Scatter plots of 100 least squares estimates (**black**) vs. lasso estimates for $\lambda = 1$ in **red** and $\lambda = 10$ in **dark red**. Lasso can aggressively shrink or zero-out coefficients, but the variance reduction is less uniform than ridge. The shift from the true (blue star) may or may not be worth it.

Lasso: Regularization enables variable selection

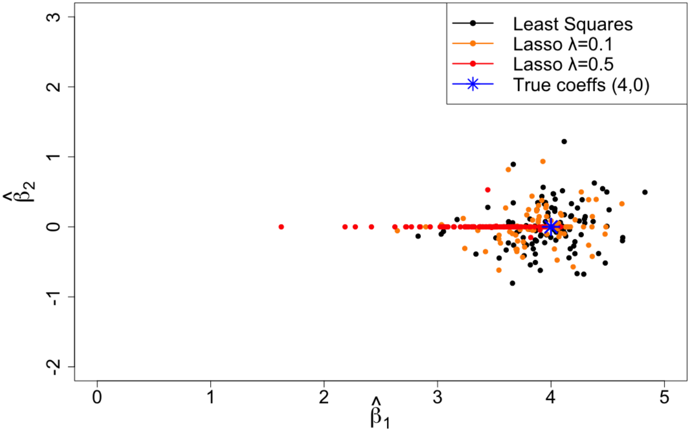

Figure: Scatter plots of 100 least squares estimates (**black**) vs. lasso estimates for $\lambda = 0.1$ in **orange** and $\lambda = 0.5$ in **red**. If the true $\beta_2 = 0$ (blue star), lasso can correctly select the significant variable (X_1), while suppressing noise and driving estimates to zero for X_2, thereby capturing the "sparse" true associations.

Lasso: Illustration with 1D example

In the setting with $n = p = 1$ without intercept, ridge solves for $\lambda \geq 0$:

$$\hat{\beta}^R_\lambda \in \arg\min \left\{ (y - x\beta)^2 + \lambda|\beta| \right\}$$

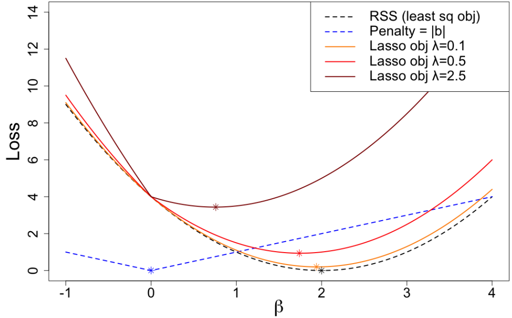

Figure: As λ grows, $\hat{\beta}^\ell_\lambda$ shrinks more aggressively; small $|y|$ can yield $\beta = 0$ ($y = 2$, $x = 1$ fixed).

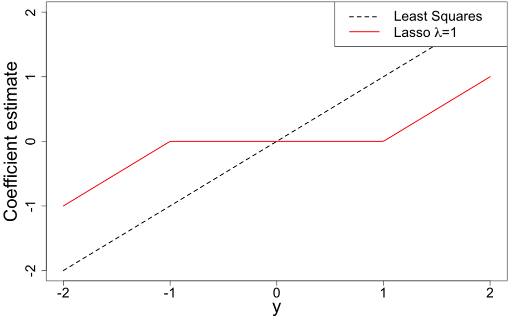

Figure: At $\lambda = 1$, $x = 1$, β hits 0 iff $|y| \leq 1$. This "thresholding" property underlies variable selection.

(Optional[1]) Alternative formulation: Constrained form

Ridge and lasso can be expressed as equivalent *constrained* optimization problems:

$$\text{(Ridge)} \qquad \text{minimize}_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \le s_\lambda$$

$$\text{(Lasso)} \qquad \text{minimize}_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le s'_\lambda$$

- For each $\lambda \ge 0$, there exist s_λ, s'_λ such that solving the above problems yield the same ridge/lasso regression coefficient estimates
- Geometrically: feasible region is an ℓ_2-ball for ridge or ℓ_1-ball for lasso

[1] That is, it is good to know, but its mathematical details will not be asked in the exams

The lasso prefers “spiky” solutions

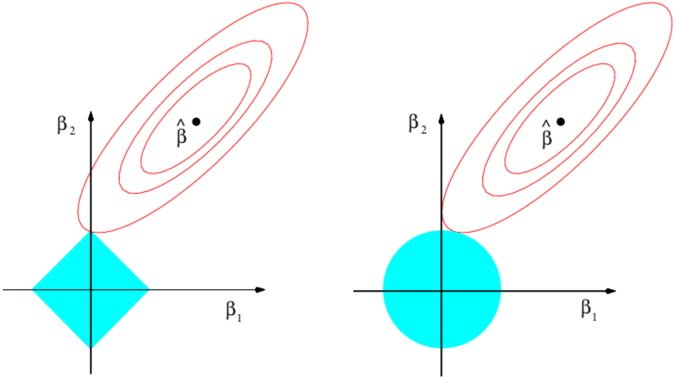

Figure: Contours of the RSS (**red** ellipses) and the feasible sets (**cyan** areas). **Left:** For lasso, the constraint $\|\beta\|_1 \leq s$ (a diamond shape) can yield corner solutions having exact zeros. **Left:** For ridge, the constraint $\|\beta\|_2^2 \leq s'$ is round, so typically yielding no exact zeros [JWHT21, Figure 6.7].

Comparison of ridge vs. lasso

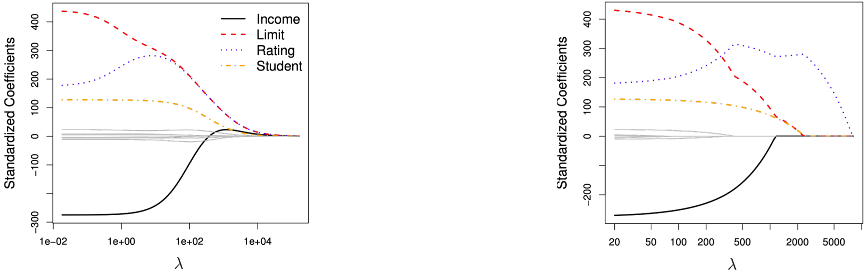

Figure: Standardized ridge (left) and lasso (right) coefficients on `Credit` dataset, plotted vs. λ [JWHT21, excerpted from Figures 6.4 & 6.6].

Ridge:

- More stable under collinearity
- Typically no exact zeros
- Often simpler closed-form solution

Lasso:

- Possibly less stable under correlated predictors
- Produces zero coefficients (variable selection)
- More interpretable if many X_j are irrelevant

Regularization: Summary

- **Why regularization?**
  - Remedy high variance or ill-posedness, especially when $p \approx n$ or $p > n$
  - Potentially yield simpler, more interpretable models (esp. lasso)
- **How?** Add a penalty
  - Ridge: $\sum_{j=1}^{p} \beta_j^2$ shrinks all β_j stably, rarely yielding exact zeros
  - Lasso: $\sum_{j=1}^{p} |\beta_j|$ can drive some β_j to 0, enabling variable selection
  - Tuning parameter λ typically selected via cross-validation
- **Ridge vs. Lasso**:
  - Ridge is stable under collinearity and has simpler closed-form solutions
  - Lasso can yield sparse solutions (some $\beta_j = 0$)
  - Neither strictly dominates: test performance depends on the data
    $\rightarrow$ usually do cross-validation to choose

Pop-up quiz #1: Regularization

Which statement is false regarding ridge and lasso?

A) **Ridge** solutions typically shrink correlated predictors together in a "group" manner.

B) **Lasso** can produce exactly zero coefficients, offering built-in variable selection.

C) Once λ is chosen by cross-validation, *ridge will always* outperform lasso in test MSE.

D) Both ridge and lasso can handle $p > n$ by imposing shrinkage or sparsity, respectively.

Answer: (C) is false.

In reality, neither ridge nor lasso *always* wins after tuning λ; their test performance is problem-dependent, so we typically compare both (often via cross-validation).

Multiple hypothesis testing: Motivation

Recall single-hypothesis testing:

- For each predictor X_j, test $H_0 : \beta_j = 0$
- Reject H_0 if $p < \alpha$ (e.g., $\alpha = 0.05$); Type I error rate $= \alpha$ for *one* test
  - **Type I** (False positive): Null is true, but we reject
  - **Type II** (False negative): Null is false, but we fail to reject

Modern data analysis often tests **many variables** (or features) simultaneously

- We want to identify which predictors are "significant" among many candidates

Examples:

- Testing thousands of genes/biomarkers for disease association
- Testing many (possibly high-dimensional $p > n$) predictors for stock price forecasting

Problem: Merely applying ordinary tests to each predictor can yield many false positives

Multiple hypothesis testing: Illustration

"Stock broker" example:

- 1,024 brokers each predict market ups/downs for 10 days
- By sheer luck, one broker might guess all 10 correctly
- Interpreting that single perfect record as "skill" ignores the 1,023 others tested

Coin-flip analogy:

- Testing *fairness* of a coin: $H_0 : p = 0.5$
- If we flip 1,024 fair coins ten times each, on average one coin is all heads[2]
- Standard test on that single coin gives p-value below 0.002

Key points:

- With many tests, extreme results can happen just by chance
- We must account for that when claiming "significance"

[2] Probability of "10 heads in a row" is $(\frac{1}{2})^{10} = \frac{1}{1024}$

Multiple hypothesis testing: Challenges

Setting:

- Suppose we have m predictors to test simultaneously
- Each test has a per-hypothesis Type I error rate $\alpha > 0$

Problem:

- With m tests, we have m chances for false positives
- Probability of ≥ 1 false rejection $\approx 1 - (1 - \alpha)^m$, which can be large as m grows
  - e.g. at $m = 20$ and $\alpha = 0.05$, we expect ≈ 1 false positive on average

How to address?

- Requiring $p < 0.05$ for each *does not* guarantee a $\leq 5\%$ chance of *any* false positive
- We need **multiple-comparison corrections** (next Lecture)
  - *Family-Wise Error Rate (FWER)* ensures probability of *any* false positive is $\leq \alpha$
  - *False Discovery Rate (FDR)* limits the *proportion* of false positives among all rejections

Pop-up quiz #2: Motivation for Multiple Testing

Which statement is <u>false</u> about multiple hypothesis testing?

A) When testing many predictors simultaneously, standard single-hypothesis $p < 0.05$ rules can lead to more than 5% chance of any false positive.

B) The probability of at least one false positive tends to *decrease* as we increase the number of tests.

C) We need some corrections to account for testing multiple hypothesese simultaneously, such as controlling the family-wise error rate or the false discovery rate.

D) Among 1,024 fair-coin flips, we expect about one coin to show 10 heads in a row purely by chance, and thus, observing 10 heads in a row may not be too surprising.

Answer: (B) is false. In fact, the chance of at least one false positive *increases* with more tests.

Wrap-up & next steps

- **Regularization:**
  - Ridge (ℓ_2 penalty) is stable under correlated predictors
  - Lasso (ℓ_1 penalty) can set some coefficients exactly to zero (variable selection)
  - Typically pick λ via cross-validation
- **Multiple hypothesis testing:**
  - Single-hypothesis framework can fail when m is large
    - Probability of at least one Type I error can be quite large
  - We need corrections for controlling false positives
- **Next time:**
  - Family-wise error rate: Bonferroni correction
  - False discovery rate control: Benjamini–Hochberg

References

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.